DATA MINING USING INTEGRATION OF CLUSTERING AND DECISION TREE

1K.Murugan, 2P.Varalakshmi, 3R.Nandha Kumar, 4S.Boobalan
1Teaching Fellow, Department of Computer Technology, Anna University
2Assistant professor, Department of Information Technology, Anna University, Anna University
3Teaching Fellow, Department of Computer Technology, Anna University
4M.E Scholar, Department of Computer Technology, Anna University

krishna.muruga@gmail.com, varanip@gmail.com, nandhakumart03@gmail.com, boobalanfriends_2009@gmail.com

Abstract- Data mining using integration of clustering and decision tree algorithm has been proposed for predicting the stock market prices. This mechanism involves studying stock price patterns in time by attempting to predict future results of a time-series by simply studying patterns in the time-series of stock prices. The goal of this project is to implement data mining in order to predict the Time-Series Stock prices by integrating clustering and Decision Tree Algorithm. The stock prices are grouped into clusters such that the data are similar to each other within a cluster. These clusters of data are then used to predict the stock Prices using decision tree.

Index Terms—Clustering, Decision Tree, Time-Series, Stock Price.

I. INTRODUCTION

Data mining is a relatively young and interdisciplinary field of computer science. It is the process of discovering new patterns from large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems. The stock market is a highly theorized domain. Our aim is to implement data mining in order to predict the Time – series stock prices using clustering with decision tree.

Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes. The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables.

For instance, prediction analysis technique can be used in sale to predict profit for the future if we consider sale as an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

In this paper, we seek to study and demonstrate the application of Clustering and Decision Tree approach for the analysis of the stock market. The goal of this project is to implement data mining in order to predict the Time-Series Stock prices by integrating clustering and Decision Tree Algorithm. The stock prices are grouped into clusters such that the data are similar to each other within a cluster. These clusters of data are then used to predict the stock prices using decision tree. The main problem with back propagation was the initial input data which decides the prediction. The problem is resolved as the similar data are grouped into clusters of data. The K – means clustering algorithm is used to split the data into clusters. The next step in the proposal is to apply decision tree algorithm to the above clusters of data that are formed.

The rest of this paper is organized as follows: Section 2 summarizes related researches. Section 3 gives a brief introduction of the proposed work. Section 4 describes the simulation results.

II. RELATED WORKS

In Stock Market, today’s price is extremely a good approximation of tomorrow’s price. Hence this system can forecast the buying and selling signs according to the prediction of future trends to stock market and provide decision- making for stock investors in the IEEE transaction “Forecasting stock prices using financial data mining and neural Network”, IEEE 2011. Multi-Layer Back Propagation algorithm is being used to predict the stock prices of the time – series stock data. The K – means clustering algorithm [2] is used to cluster the data into groups which resolve the Classification problem which was cited earlier while using back propagation. Hence the initial input is now the clusters of data which helps in predicting the data. The accuracy of the prediction [4] is being improved by clustering the data in the first stage as the similar data are grouped into clusters in the IEEE transaction, “Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations”. It is cited that the speed is relatively increased [8] with the help of using decision tree algorithm instead of back propagation. The concept of clustering with decision tree [7] makes the entire prediction process an easier step. Decision trees are commonly used in
operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Another use of decision trees is as a descriptive means for calculating conditional probabilities.

Classification generally uses supervised learning approaches by assuming a labeled training set. However, in many cases, there is no clear structure in the sampled data or training set where classes are not defined. The number of classes is unknown and the relationship of typical attributes to the classes may not seem obvious. Besides, over time, the characteristics of the class-specific pattern generating systems can change continuously. In these cases, an unsupervised system is preferred where an automatic and continuous learning and adaptation is possible. These types of techniques are generally referred as clustering in which class labels will not be assigned and the system must determine the natural „partitions“ or „clusters“ of the data.

Cluster analysis is the automatic identification of groups of similar objects or patterns. For example, if a set of data denoted by x, is very similar to a few other sets of data, we may intuitively tend to group x and these sets of data into a natural

In this paper, we focus on C4.5 [12] which is one of decision tree generators using top-down approach. It uses the information gain ratio as the splitting criterion for each internal node. Like other similar top-down approaches, C4.5 uses a greedy searching strategy with looking one step ahead to find the way of splitting instance space, so it often suffers from being trapped at local optimum and therefore performs poorly in dealing with some hard classification tasks, in which training data are described by high dimensional attribute vectors and the concept to be learned is complex. There have been proposed several non-greedy search approaches, which choose a splitting model automatically by observing the structure of data, so the splitting models are different nodes to nodes. Construction of optimal or near-optimal decision trees using a two-stage approach has been attempted by many authors. In the first stage, a sufficient partitioning is induced using any reasonable greedy method. In the second stage, the tree is refined to be as close to optimal as possible. Another temptation of generating non-greedy decision tree is to use genetic method.

This paper proposes a new algorithm for generating a tree classifier that is based on C4.5 and an instance-based learning. The new approach is called CC4.5. Unlike other decision tree generation methods, CC4.5 uses clustering as a pre-processing procedure and the decision tree is generated according to the result of clustering. C4.5 generates a decision tree using the standard TDIDT (top-down induction of decision trees) approach, recursively partitioning the instance space into smaller subspaces, based on the value of a selected attribute. It begins with a set of instances, called training instances, already divided into classes. Each instance is described in terms of a set of attributes, which can be numerical or symbolic. The overall approach uses a greedy search strategy to choose the attribute that divides the instances best into their classes. This process is applied recursively to each partitioned subset, with the procedure ending when all instances in the current subset belong to the same class; C4.5 uses an information gain ratio for measuring how well an attribute can divide the instances into their classes.

In the paper [10], “Integrating Decision Tree and Spatial Cluster Analysis for Landslide Susceptibility Zonation”, a concept called CART (Classification and Regression Tree) is implemented. The main idea in CART is to partition the dataset into homogeneous subgroups with respect to the same class. The complex data structure can be represented conveniently by a tree structure in which an internal node denotes a best split predictor variable, the branches of a node denote the criteria value of the split variable, and a leaf denote the final response class. In the tree structure, the paths from the root node (top node) to leaf (terminal node) show the decision rules that maximize the distinction among the classes and minimize the diversity in each class.

III. PROPOSED WORK

To implement the stock prices are grouped into clusters such that the data are similar to each other within a cluster. These clusters of data are then used to predict the stock Prices using decision tree. The main problem with back propagation was the initial input data which decides the prediction. The problem is resolved as the similar data are grouped into clusters of data. The K – means clustering algorithm is used to split the data into clusters. The next step in the proposal is to apply decision tree algorithm to the above clusters of data that are formed. The proposed architecture is designed to overcome the disadvantages of the existing work such that the back propagation algorithm. The C4.5 decision tree algorithm is used for building the decision tree. The input to the algorithm is the collection of training data cases consisting of the inter day stock data

PROPOSED SYSTEM ARCHITECTURE

![Fig 1 System Architecture](image)

In Fig 1, the time-series stock price is given as the input to
the K - means clustering algorithm. By applying this K - means clustering algorithm, the stock prices are grouped into different clusters such that the data within the cluster are similar to each other. Here the cluster centers are initialized with those k clusters. After initializing the cluster centers, perform partitioning by assigning or reassigning all data objects to their closest cluster center. Compute new cluster centers as mean value of the objects in each clusters until no change in cluster center calculation. This cluster partitioning is done repetitively until there is no change in the cluster center calculation. These clusters of data are then used to predict the stock prices using the decision tree algorithm. The inter day stock data of the companies are taken as the training sets upon which the decision tree is built. The C 4.5 decision tree algorithm is used for building the decision tree. The input to the algorithm is the collection of training data cases consisting of the inter day stock data. The goal is to predict the outcome using the training sets of data which is done by C 4.5 algorithm. Finally, the integration of clustering and decision tree gives the predicted result.

In data mining, k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. This results into a partitioning of the data space into Voronoi cells. The problem is computationally difficult (NP-hard), however there are efficient heuristic algorithms that are commonly employed and converge fast to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data, however k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

**PROPOSED ALGORITHM**

**K -MEANS CLUSTERING KMEANS(X, CI) → (C, P)**

**REPEAT**

*Cprevious ← Cl;*

FOR all i ∈ [1, N] DO

Generate new optimal partitions

\[ p(i) ← \arg \min d(x_i, c_j); \]

1 ≤ j ≤ k

FOR all j ∈ [1, k] DO

Generate optimal centroids

\[ c_j ← \text{Average of } x_i, \text{ whose } p(i) = j; \]

UNTIL C = Cprevious WHERE

\[ X: \text{a set of } N \text{ data vectors} \]

CI: initialized k cluster centroids

\[ C: \text{the cluster centroids of } k\text{-clustering} \]

\[ P = \{p(i) \mid i = 1, \ldots, N\} \text{ is the cluster} \]

**label of X**

**X : Data set**

K : number of clusters

**Fig 2 K means Clustering Algorithm**

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set \( S = \{s_1, s_2, \ldots \} \) of already classified samples. Each sample \( S_i = x_1, x_2, \ldots x_n \) is a vector where \( x_1, x_2 \ldots \) represent attributes or features of the sample. The training data is augmented with a vector \( C = c_1, c_2 \ldots \) where \( c_1, c_2 \ldots \) represent the class to which each sample belongs.

At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one continuous and discrete attributes. In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.

**C4.5 ALGORITHM**

Form tree (T)

\[
\begin{align*}
\text{Compute class frequency (T)} & \\
& \text{If One class} \\
& \quad \text{return a leaf; } \\
& \quad \text{create a decision node N; } \\
& \text{For each attribute A} \\
& \quad \text{N.test = max(compute gain(A))} \\
& \text{for each } T^* \text{ in the splitting of T} \\
& \quad \text{if } T^* \text{ is empty} \\
& \quad \quad \text{child of N is a leaf; } \\
& \quad \quad \text{else} \\
& \quad \quad \text{child of N = Form tree(T*)} \\
& \quad \text{return N}
\end{align*}
\]
B. Simulation Results

Fig 4 K clusters formed using K means clustering

The figure 4 explains the clusters formed by the K means Clustering Algorithm. Stock Prices of various industries are given as the dataset to the algorithm. Then K clusters are identified in the dataset and each similar data groups are placed in a separate cluster. These clusters of data are then used to predict the stock Prices using decision tree. Here the first two columns represent stock prices and market capacity of various industries. Then the K means Clustering algorithm is applied to both these attributes and the stock prices are grouped into clusters such that the data within a cluster are similar to each other. Thus the third column in the fig 4.1 represents the cluster number to which the data belongs.

Fig 5 Decision Tree for cluster (7)

The figure 5 represents the decision tree formed by applying the C4.5 algorithm to the cluster 7. The decision tree is formed by generating various rules using C4.5 algorithm. The attributes that are considered for constructing the decision tree are market cap, stock price, firm value, correlation with market, hi-low risk measure, trailing net income, current invested capital, net profit margin. In the first level of the tree, if the gain ratio of the attribute, trailing income” is greater than 1.34, then the stock price is predicted to rise (prediction-1). Otherwise, check the gain ratio of the next attribute 5 „stock price”. If the gain ratio of stock price is greater than 0.5, then it is predicted to rise. Else if the gain ratio of „correlation with market” is greater than 16.73, the stock price is predicted to rise. In the fourth level of the decision tree, the gain ratio of the attribute, net profit” is checked. If it is greater than 3.57, stock price tends to rise else it will fall. Then in the fifth level, if the gain ratio of firm value is greater than 3.34, then the corresponding stock price will rise. Similarly the gain ratios of current capital, market cap are checked and the prediction is done.

IV. CONCLUSION AND FUTURE WORK

Classification of data is a major advantage and their performance in large data sets is useful in the prediction. Thus the prediction of stock prices is done by integrating the clustering and decision tree algorithm.

This paper describes a cluster analysis based on decision theory. The proposal uses a loss function to construct the quality index. Therefore, the cluster quality is evaluated by considering the total risk of categorizing all the objects. Such a decision-theoretic representation of cluster quality may be more useful in business-oriented data mining than traditional geometry-based cluster quality measures. A real advantage of the decision-theoretic cluster validity measure is its ability to include monetary considerations in evaluating a clustering scheme. We can also extend it to evaluating other clustering algorithms such as fuzzy clustering. Such a cluster validity measure can be useful in further theoretical development in clustering. Results of such development will be reported in the future.

V. REFERENCES


[5] Bingyu Cao, Deping Zhan, Xianbin Wu,”Application of SVM in Financial Research”, International Joint Conference, pp 507-
511, 2009.


